Overview of the PerAnsSumm 2025 Shared Task on Perspective-aware Healthcare Answer Summarization

Siddhant Agarwal[1], **Md. Shad Akhtar**[2], **Shweta Yadav**[1],
[1]University of Illinois at Chicago, [2]IIIT Delhi
{sagarw38, shwetay}@uic.edu, shad.akhtar@iiitd.ac.in

Abstract

This paper presents an overview of the Perspective-aware Answer Summarization (PerAnsSumm) Shared Task on summarizing healthcare answers in Community Question Answering forums hosted at the CL4Health Workshop at NAACL 2025. In this shared task, we approach healthcare answer summarization with two subtasks: (a) perspective span identification and classification and (b) perspective-based answer summarization (summaries focused on one of the perspective classes). We defined a benchmarking setup for the comprehensive evaluation of predicted spans and generated summaries. We encouraged participants to explore novel solutions to the proposed problem and received high interest in the task with 23 participating teams and 155 submissions. This paper describes the task objectives, the dataset, the evaluation metrics and our findings. We share the results of the novel approaches adopted by task participants, especially emphasizing the applicability of Large Language Models in this perspective-based answer summarization task.

1 Introduction

Community Question Answering (CQA) forums such as Yahoo! Answers, Reddit, and Quora have transformed how people access information, especially with the rise of the internet. These sources facilitate the spread of information and knowledge across geographical boundaries and connect people with wide-ranging expertise and experiences. It is therefore no surprise that users of these forums discuss a broad range of topics, including healthcare concerns. However, within these forums, users often struggle to find relevant and reliable information given the plethora of answers. Further, these forums contain answers from users with a multitude of perspectives, such as their personal experiences or subject knowledge, which may or may not be relevant to what another user seeks. To this end, Naik et al. (2024) proposed the perspective-aware healthcare answer summarization task for CQA forums.

As seen in Figure 1, users' questions often receive answers from other users of CQA forums that contain a multitude of perspectives. For example, a user provides both a suggestion ("*try a diet with low fat and very low saturated fats*") and their personal experience ("*I've had the surgery and it really isn't a big deal*") in their answer. While such diverse insights can be valuable, they can also be overwhelming for users seeking specific information. Therefore, it is important to identify such perspective spans and provide a concise perspective-based summary of all answers (as shown in Figure 1). This allows users to obtain information relevant to their situation and assists them in making informed decisions.

The investigation of novel approaches for the task of CQA forum answer summarization has been limited with recent works being primarily reliant on Pre-trained Language Models (Naik et al., 2024) such as Flan-T5, leaving the utility of Large Language Models unexplored for the most part. Further, the majority of previous work has been limited by small dataset sizes (Bhattacharya et al., 2022; Chaturvedi et al., 2024) and the lack of a uniform benchmark. This work aims to fill this research gap by providing an accessible resource to researchers for developing and comparing novel techniques for perspective-aware healthcare answer summarization.

The PerAnsSumm 2025 Shared Task focuses on investigating novel solutions in the perspective-aware summarization of healthcare answers in CQA forums. This work aims to be a meaningful step forward in spearheading research in this direction and investigating the utility of recent advances in Natural Language Processing, such as the rise of Large Language Models (LLMs) in their application to the biomedical summarization domain.

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 445–455
May 4, 2025 ©2025 Association for Computational Linguistics

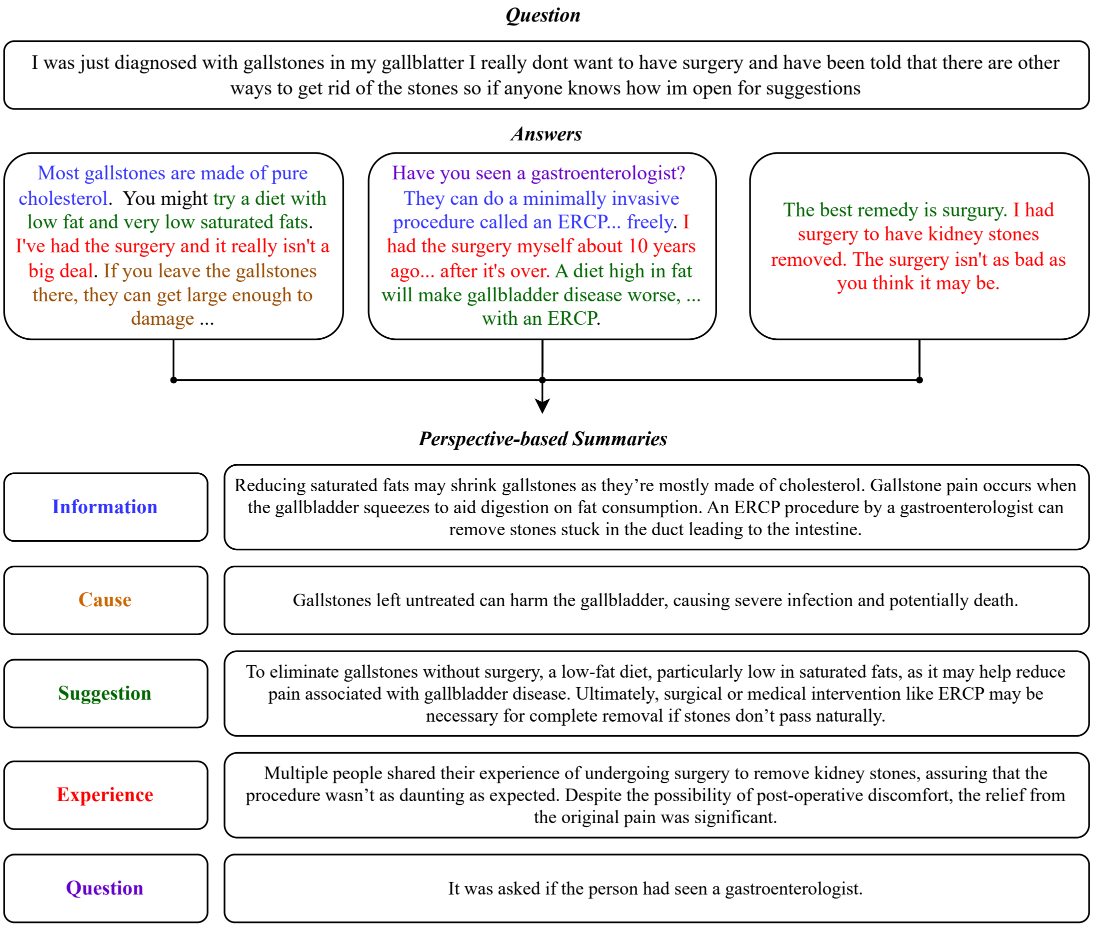

Figure 1: A description of the PerAnsSumm task with inputs and expected output. Colored spans in answers correspond to spans of different perspectives. The spans are utilized to generate a perspective-based summary for each class.

In this work, we present the findings of the PerAnsSumm 2025 Shared Task, hosted by the CL4Health Workshop at NAACL 2025. The shared task garnered significant interest, with 100 registered participants on the CodaBench[1] platform, with 23 teams participating and submitting a total of 155 valid submissions. The remainder of this paper describes key details of the shared task along with our findings and brief descriptions of the participating systems.

2 Task Description

The shared task involved two sub-tasks, (A) Span-Identification and Classification and (B) Summary Generation. These two sub-tasks aimed to capture the different ways in which a user may interact with a Community Question-Answering Forum when filtering based on the five defined perspectives – *'Information'*, *'Cause'*, *'Suggestion'*, *'Experience'* and *'Question'*.

TASK A – Perspective Span Identification and Classification. In this task, the participants were required to identify and accurately classify spans of text in the community answers of CQA threads according to the relevant perspective. For example, as shown in Figure 1: `Information` - '*gallstones are made of pure cholesterol*', `Experience` - '*I had the surgery myself about 10 years ago*', `Question` - '*Have you seen a gastroenterologist*'.

TASK B – Perspective-based Summarization. In this task, participants were required to provide a summary of all texts pertaining to the relevant perspective class. This may be looked at as a summary of the identified perspective-based spans or as a perspective-based summary of the answers in the CQA thread. For example, as shown in Figure 1: `Cause` - '*Gallstones left untreated can harm the gallbladder, causing severe infection and poten-*

[1] Available for continued access at: `https://www.codabench.org/competitions/4312/`

	Size	Information	Cause	Suggestion	Question	Experience
Train	2533	4823/1961	646/342	646/342	325/249	1439/845
Validation	317	643/246	108/49	549/208	42/32	170/108
Test (Seen)	317	631/242	81/45	499/188	44/31	181/100
Test (Unseen)	50	153/43	47/14	198/47	35/18	92/37

Table 1: Dataset Statistics describing the perspective-specific span count/summaries count in the split

tially death.'

Both tasks combined to address the underlying challenge of providing users with relevant content that is specific to their needs, and hence, allowing them to make informed decisions. We proposed these tasks as complementary, as identifying relevant perspective-specific spans allowed for improvements in the summarization task. However, the participating teams were given the option to participate in each task individually if they preferred.

3 Dataset

For this task, we utilized the PUMA dataset (Naik et al., 2024), containing 3167 total questions and 9987 answers. The dataset is divided into training, validation, and testing sets with detailed class-wise statistics given in Table 1. The PUMA dataset was developed using samples from the **L6 - Yahoo! Answers CQA** dataset [2] filtered on the Healthcare category. These samples were annotated by analyzing all answers for potential perspective labels and manually writing a perspective-based summary that is a concise representation of all perspective spans. As a result of this annotation, the dataset contained text spans in each answer, along with a perspective-based answer summary for each identified perspective class label for a question sample.

Naik et al. (2024) identified five perspective classes that correspond to the different ways in which users respond to questions on CQA forums. These perspectives were given as follows:

1. **Cause:** It underlines the potential cause of a medical phenomenon or a symptom. It answers the *Why* regarding a specific observation, offering insights to identify the root cause.

2. **Suggestion:** It encapsulates strategies, recommendations, or potential courses of action towards management or resolution of a health condition.

3. **Experience:** It covers first-hand experiences, observations, insights, or opinions derived from treatment or medication related to a particular problem.

4. **Question:** It consists of interrogative phrases, follow-up questions and rhetorical questions that are sought to better understand the context. They typically start with phrases like *Why, What, Do, How,* and *Did* etc, and end in a question mark.

5. **Information:** It encompasses segments that offer factual knowledge or information considering the given query. These segments provide comprehensive details on diagnoses, symptoms, or general information on a medical condition.

Through our utilization of this dataset, we hope to enable researchers to develop models which are capable of generating perspective-guided summaries for CQA answer forums. This would in turn enable users to make informed decisions when accessing CQA forums.

Since the original PUMA dataset was available to researchers upon request, we further annotated 50 samples as a new test set for the PerAnsSumm shared task. We followed the annotation guidelines as laid out by Naik et al. (2024) to identify relevant spans for each perspective class and manually created summaries for the identified perspectives. Submissions by the participants to the PerAnsSumm 2025 Shared Task were evaluated on this set of 50 newly annotated and unreleased samples.

4 Evaluation

In this section, we provide details about the evaluation metrics used for each of the two sub-tasks in the PerAnsSumm 2025 shared task.

Task A We evaluated submissions on 3 criteria - Classification (Macro F1 and Weighted F-1), Strict-matching (Precision, Recall and F-1), Proportional-matching (Precision, Recall and F-1). The overall score for task A combined these 3 criteria as it is the average of the classification-weighted F-1 score, the Strict-matching F-1 score and the Proportional-matching F-1 score. Classification metrics were based on framing the problem as a sample-level multi-label classification problem. Strict matching was defined as follows:

$$P = \frac{|\text{CorrectSpans}|}{|\text{PredictedSpans}|},$$

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11

$$R = \frac{|\text{CorrectSpans}|}{|\text{GoldSpans}|},$$

$$F_1 = \frac{2 \times P \times R}{P + R},$$

Proportional-matching was defined as follows:

$$P = \frac{\sum len(\text{MaximumOverlappingSpan})}{\sum len(\text{PredictedSpan})},$$

$$R = \frac{\sum len(\text{MaximumOverlappingSpan})}{\sum len(\text{GoldSpan})},$$

$$F_1 = \frac{2 \times P \times R}{P + R},$$

where MaximumOverlappingSpan refers to the sub-span of a predicted span that had the maximum overlap with each of the gold spans.

Task B Submissions were evaluated based on two criteria using multiple automatic metrics to assess both the relevance and the factuality of the generated summaries. These criteria were as follows:

1. *Relevance* - ROUGE-1,2 and L (Lin, 2004), BertScore (Zhang et al., 2020b), METEOR (Banerjee and Lavie, 2005) and BLEU (Papineni et al., 2002).

2. *Factuality* - AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2022).

The overall score across both tasks was computed as an average of the Task A scores, the Task B Relevance scores, and the Task B Factuality scores. This was used in computing the final leaderboard positions. Implementation and hyperparameters used for all automatic evaluations were made available [3] to the participants before the evaluation stage.

5 Task Results

Table 2 presents the final leaderboard for the shared task based on the best performing submission of each team, according to the defined evaluation metrics. Task-wise results are given in Table 3 and 4.

In this section, we describe our findings and key results from the submissions.

[3]Made available through a GitHub repository: https://github.com/PerAnsSumm/Evaluation

$\star$	**Team**	**LLMs?**	**Score**
1	WisPerMed	✓	45.71
2	YALENLP	✓	45.48
3	Team_ABC	✓	45.26
4	AICOE	✓	44.95
5	KHU_LDI	✓	44.92
6	LTRC-IIITH	✓	43.95
7	MNLP	✓	43.21
8	Team Airi	✓	42.38
9	DataHacks	✓	42.03
10	UTSA-NLP	✗	41.12
11	HSE NLP	✓	40.81
12	MediFact	✓	40.77
13	NU-WAVE	✓	40.46
14	Roux-lette	✓	39.96
15	Manchester Bees	✓	39.94
16	Abdelmalak	✓	39.07
17	Team_UMB	✗	38.24
18	massU	✗	38.15
19	RVK_Med	✗	37.50
20	TrofimovaMC	✗	36.98
21	TeamENSAK	✓	36.41
22	CaresAI	✓	34.05
23	LMU*	✓	17.26

Table 2: Final leaderboard for the PerAnsSumm 2025 Shared Task in order of average performance over the two sub-tasks. $\star$ denotes the rank column. Combined Average is the average of the average Task A and average Task B scores. * denotes the team participates in Task B only.

LLM usage As a part of the submission process, we asked participants to self-disclose the use of LLMs in their modeling approaches. Out of 23 participating teams, 18 teams disclose the use of LLMs in some capacity, with all of the top 10 teams utilizing LLMs. This highlights the growing prevalence and importance of LLMs in summarization and other NLP tasks. The growing trend of LLM utilization is highlighted further when compared to a similar task related to summarization in the biomedical domain, BioLaySumm 2024 (Goldsack et al., 2024), where only 18 of the 52 participating teams utilized LLMs. The rapidly evolving landscape of LLM research and its applications in the biomedical domain need careful evaluation, especially given the sensitivity of biomedical data and the related real-life implications. At the same time, we find this usage of LLMs as a positive signal of participants exploring novel techniques.

*	⋆	Team	Classification		Strict-matching			Proportional-matching			Avg
			macro	*weigh.*	*Prec.*	*Recall*	*F1*	*Prec.*	*Recall*	*F1*	
1	3	Team_ABC	86.97	91.73	22.05	27.81	**24.60**	62.15	80.29	70.06	**62.13**
2	7	MNLP	85.24	90.61	13.76	27.24	18.29	65.80	84.06	73.82	60.90
3	4	AICOE	86.56	91.40	17.65	27.43	21.48	65.97	71.59	68.66	60.52
4	2	YALENLP	84.39	89.02	15.71	28.57	20.27	63.72	82.18	71.78	60.36
5	6	LTRC-IIITH	**90.33**	**92.39**	19.15	22.29	20.60	67.74	68.33	68.03	60.34
6	12	MediFact	83.61	88.87	13.83	31.43	19.21	62.22	**84.93**	71.82	59.97
7	1	WisPerMed	87.75	92.11	17.26	23.05	19.74	62.36	73.80	67.60	59.82
8	16	Abdelmalak	88.59	91.30	8.53	15.81	11.08	**70.21**	81.74	**75.54**	59.31
9	5	KHU_LDI	79.09	86.18	18.68	**30.10**	23.05	57.16	81.84	67.31	58.85
10	13	NU-WAVE	81.24	87.19	20.48	22.86	21.60	57.02	72.26	63.74	57.51
11	14	Roux-lette	81.24	87.19	20.48	22.86	21.60	57.02	72.26	63.74	57.51
12	15	Manchester Bees	82.68	87.69	**22.67**	19.43	20.92	55.03	70.36	61.76	56.79
13	10	UTSA-NLP	73.59	84.26	16.87	18.67	17.72	59.66	67.64	63.40	55.13
14	11	HSE NLP	80.73	87.86	14.75	18.86	16.56	66.66	54.21	59.79	54.74
15	9	DataHacks	86.35	90.44	15.99	13.52	14.65	51.49	66.78	58.15	54.41
16	8	Team Airi	84.67	88.67	19.94	12.76	15.56	49.13	61.67	54.69	52.98
17	19	RVK_Med	89.84	92.07	0.19	0.19	0.19	58.01	72.05	64.27	52.18
18	18	massU	83.16	88.54	14.29	11.43	12.70	50.85	48.30	49.54	50.26
19	17	Team_UMB	82.91	88.26	12.66	11.43	12.01	52.32	48.77	50.48	50.25
20	20	TrofimovaMC	77.00	85.79	7.28	9.52	8.25	58.14	46.30	51.55	48.53
21	21	TeamENSAK	80.69	84.94	1.69	2.10	1.87	58.23	46.02	51.41	46.08
22	22	CaresAI	74.64	83.02	7.37	8.00	7.67	47.54	36.51	41.31	44.00

Table 3: Leaderboard for Task A of the PerAnsSumm 2025 Shared Task in order of average performance. ⋆ denotes the overall shared task rank column. ∗ denotes Task A rank column. Classification scores are F1 scores. Average for Task A is calculated as the average of classification-weighted F1, Strict-matching F1, and Proportional-matching F1.

Comparing Task A and Task B performance We find that teams that perform well in Task A, which covers identification and classification, also tend to perform comparatively better in Task B, perspective-based summarization. It is observed that teams often utilize substantially different methods for both the tasks, with greater reliance on smaller pre-trained language models in the span identification task compared to the summarization task.

In-context learning as the new normal An interesting observation from the submissions is the reliance on novel in-context learning based approaches through innovative prompting strategies. Participants prefer inferencing on pre-trained large language models, utilizing their vast training knowledge as compared to fine-tuning models specifically for the task. This reliance is representative of the current shift in the NLP landscape from a pre-train and fine-tune to a pre-train and inference paradigm. This calls for the further development of models trained specifically on specialized domains, such as healthcare to advance research and boost model capabilities in these specialized areas.

6 Submissions

The PerAnsSumm 2025 shared task attracted submissions from 23 participating teams who made a combined total of 155 valid submissions that were evaluated by the task organizers. Out of these teams, 22 teams participated in both Task A and Task B, while 1 team participated in only Task B. Out of the 23 participating teams, 12 teams submitted system papers. Brief summaries of the approaches taken by these teams are described in this section. We also describe the baseline provided to participants as a starter code.

Starter Kit: We utilized the PLASMA model (Naik et al., 2024) as a strong starting point to the participants. This modeling approach showed promising results in the perspective-based answer summarization task (Task B). It utilized a perspective-conditioned prompt that is generated following a defined prompt template. Subsequently, the prompt was fed to the Flan-T5 model (Chung et al., 2022) with a prefix tuner to generate the summary. An energy-driven loss function was incorporated along with the standard cross-entropy (CE) loss to enforce the perspective attributes in the generated summary. This model represented the current state

∗	⋆	Team	Relevance							Factuality			Avg
			R-1	R-2	R-L	BS	MT	BL	Avg	AS	SC	Avg	
1	1	WisPerMed	45.15	22.10	41.02	89.91	40.95	13.47	42.10	40.85	**29.58**	35.21	**38.66**
2	2	YALENLP	**46.90**	**23.14**	**42.87**	88.28	**44.54**	**15.71**	**43.57**	37.94	27.07	32.50	38.04
3	5	KHU_LDI	45.48	20.44	40.31	**90.12**	39.50	14.13	41.66	42.00	26.53	34.27	37.96
4	4	AICOE	43.45	18.69	38.78	86.58	38.44	11.24	39.53	42.60	27.01	34.80	37.17
5	8	Team Airi	38.42	18.68	35.19	76.80	33.93	13.96	36.16	47.28	28.72	38.00	37.08
6	3	Team_ABC	40.01	16.49	35.78	84.06	31.87	10.60	36.47	46.01	28.34	37.17	36.82
7	9	DataHacks	37.08	16.83	33.65	77.62	33.91	11.16	35.04	44.27	28.99	36.63	35.84
8	6	LTR-IIITH	39.46	17.41	35.12	83.11	34.07	13.38	37.09	41.84	27.01	34.42	35.76
9	7	MNLP	40.22	16.39	36.08	84.93	38.85	10.70	37.86	36.17	25.53	30.85	34.36
10	10	UTSA-NLP	34.38	12.61	30.53	76.87	31.16	10.24	32.63	45.03	26.20	35.62	34.12
11	11	HSE NLP	30.84	9.61	26.03	83.36	20.62	3.81	29.05	**51.50**	25.78	**38.64**	33.84
12	17	Team_UMB	36.02	15.46	32.78	82.32	33.93	9.58	35.02	33.26	25.62	29.44	32.23
13	18	massU	36.27	15.84	33.32	82.26	34.55	9.44	35.28	32.03	25.77	28.90	32.09
14	13	NU-WAVE	38.44	16.67	33.95	82.74	33.35	12.41	36.26	32.16	23.06	27.61	31.93
15	21	TeamENSAK	30.67	12.84	27.67	69.74	25.48	11.19	29.60	41.10	25.99	33.54	31.57
16	15	Manchester Bees	29.23	9.11	24.54	77.34	21.18	4.04	27.57	47.75	23.16	35.45	31.51
17	20	TrofimovaMC	28.76	9.12	23.85	81.77	19.31	2.13	27.49	46.79	23.04	34.91	31.20
18	14	Roux-lette	37.37	15.42	32.67	82.52	32.84	11.22	35.34	31.15	22.88	27.02	31.18
19	12	MediFact	34.85	14.75	32.12	83.36	31.20	10.78	34.51	31.21	24.48	27.84	31.18
20	19	RVK_Med	30.11	11.40	27.05	81.96	26.87	8.86	31.04	33.87	24.67	29.27	30.16
21	22	CaresAI	28.00	8.45	24.31	85.00	22.06	6.12	28.99	33.14	25.21	29.17	29.08
22	16	Abdelmalak	31.32	11.30	25.56	79.88	23.40	6.34	29.63	33.84	22.70	28.27	28.95
23	23	LMU	21.48	9.05	19.42	53.51	20.32	5.95	21.62	35.64	24.71	30.17	25.90

Table 4: Leaderboard for Task B of the PerAnsSumm 2025 Shared Task in order of average performance. ⋆ denotes the overall shared task rank column. ∗ denotes Task B rank column. Task B metrics - R-1 (ROUGE-1), R-2 (ROUGE-2), R-L (ROUGE-L), BS (BertScore), MT (METEOR), BL (BLEU), AS (AlignScore), SC (SummaC). All metrics are F-1 scores wherever relevant. Average column is the average of the average Task B Relevance and Task B factuality average scores.

Team	Coherence	Consistency	Fluency	Relevance	Coverage
WisPerMed	4.40	4.40	4.47	4.00	4.07
YALENLP	**4.73**	4.53	4.60	4.20	4.40
Team_ABC	4.07	3.93	4.33	3.73	3.60
AICOE	4.27	4.00	4.40	3.73	3.80
KHU_LDI	4.53	**4.67**	**4.67**	**4.33**	**4.53**

Table 5: Human Analysis of 15 generated summaries for the top 5 ranking teams

of the art for the task of perspective-based answer summarization, and the source code for this model is provided to the participants in the starter kit as a part of the Shared Task.

WisPerMed Pakull et al. (2025) leveraged DeepSeek-R1 (DeepSeek-AI, 2025) in a zero-shot setting with structured prompting for Task A. They designed a detailed system prompt instructing the model to extract spans according to the given perspectives without modifying the original content. They instruct the model to return structured output for consistency and easy parsing. For Task B, they utilized two step pipeline with sequence classification and instruction tuning of the Mistral-7B model (Jiang et al., 2023). In the first step of this pipeline, they built a labeled answer dataset by associating the spans with their corresponding classes and using the Mistral model as a sequence classifier. In the next step, the perspective-specific subset of answers was used to generate perspective-aware summaries. The team achieved first rank on the leaderboard based on the average over Task A and Task B metrics, and also lead performance in Task B. Their approach exhibits close to peak performance across all aspects of the two tasks leading to a high overall rank compared to other teams' approaches which ace one set of metrics while falling behind on the overall task.

YALENLP Jang et al. (2025) utilized the zero-shot capabilities of GPT-4o (OpenAI et al., 2024b) for both Task A and Task B. They inference on GPT-4o without fine-tuning and rely upon the effectiveness of GPT4o to capture the diverse medical perspectives in CQA forums with promising results. They highlight that the generalizability of the GPT4o model allows for robust in-context learning and even surpasses few-shot configurations. They also utilized a Mixture-of-Agents (Wang et al., 2024) setup to enhance system performance through ensembling multiple open-source models, allowing them to compensate for the weaknesses of individual models. They exploited an intermediate verification layer to refine predictions and mitigate hallucinations. They achieved second rank on the task leaderboard with the best score Task B relevance metrics.

AICOE R et al. (2025) utilized a pipeline with a combination of two closed-source LLMs inferenced for both Task A and Task B. For Task A, they employed the OpenAI O1 (OpenAI et al., 2024a) and the Google Gemini-2.0 Flash models. The spans predicted by both these models are merged with a preference given to the Gemini-2.0 model based on an empirical review of performance. They then used these predicted spans as an additional input for Task B summarization using the Gemini 2.0 Flash model. They also highlight their experiments with fine-tuned open-source LLMs.

LTRC-IIITH Marimuthu and Krishnamurthy (2025) fine-tuned BERT-large (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019) models for span identification in the standard IO annotation format. They demonstrate the robustness of a fine-tuned RoBERTa model with the highest classification-weighted F-1 score for Task A. For Task B, they fine-tune BART-large (Lewis et al., 2020) and Pegasus-large (Zhang et al., 2019) models with an MLM (Masked-Language Modeling) objective for the BART model.

MNLP Lee et al. (2025) followed a two-stage Classifier-Refiner Architecture (CRA) to improve the classification of user-generated health responses in CQA forums. In the first stage, a classifier segments responses into self-contained snippets and assigns one of five perspective classes. If the classifier was uncertain, a refiner was triggered to reassess the classification using retrieval-augmented generation (RAG). The refiner retrieved the two most similar training examples based on all-MiniLM-L6-v2 sentence similarity and incorporated them as few-shot examples to enhance classification reliability. Additionally, they employed instruction-based prompting, tone definitions, and Chain-of-Thought (CoT) reasoning to guide the model's decisions and improve interpretability.

DataHacks Nawander and Reddy (2025) utilized the Mistral 7B (Jiang et al., 2023) model as their backbone for fine-tuning with LORA adapters. The same configuration of fine-tuning an LLM with Low-Rank Adaptation (Hu et al., 2022) was used for both tasks. They perform prompt engineering to restructure the input into the distinct sections of Question, Context, and Answer, allowing the model to better interpret details and leading to an observed improvement in model performance.

Team_UMB Qi et al. (2025) employed an ensemble learning approach combining multiple transformer models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021)) through weighted averaging for Task A. For Task B, they developed a suite of prompting techniques to leverage a pre-trained LLM (Llama-3 (Grattafiori et al., 2024)). Specifically, they used chain-of-thought (CoT) techniques with integrated keyphrases and additional guidance information. To optimize these prompts, they applied the DSPy framework with a designed downstream evaluation metric aimed at balancing relevance and factuality. Using the 0-shot MIPRO optimizer within DSPy, they iteratively optimized prompts to enhance summary generation capabilities. Furthermore, they demonstrated that incorporating supervised fine-tuning improved the quality of generated summaries.

MediFact Saeed (2025) presented a three-stage hybrid pipeline for Task A consisting of weak supervision with Snorkel, supervised learning with SVM and zero-shot classification using transformers. The transformer model was deployed in case of uncertainty in the predictions of the previous stages. For Task B, Saeed (2025) proposed an approach consisting of extractive summarization using the BART (Lewis et al., 2020) model and abstractive refinement using Pegasus (Zhang et al., 2020a).

Roux-lette Antony et al. (2025) used an LLM-based approach with semantic similarity-guided in-context learning (ICL). For Task A, they queried the Qwen-Turbo LLM (Qwen et al., 2025) by prompting it with 20 In-Context Learning samples selected from the training data using NVIDIA NV-Embed-v2 (Lee et al., 2024) text embedding model to obtain spans for each perspective. These spans were then processed through a matching pipeline that attempted exact matches first, followed by case-insensitive and fuzzy matching if needed. For Task B, they used a similar ICL-based approach, selecting relevant examples based on semantic similarity between the input text and training examples. The LLM leveraged these examples, along with the extracted spans from Task A, to generate perspective-aware summaries. The most effective prompt asked the model to replicate the annotation patterns observed in the ICL samples, ensuring that the summaries maintained alignment with human annotations.

Manchester Bees Romero et al. (2025) proposed an approach with Iterative Self-Prompting (ISP) with the closed source LLMs Claude and o1. They used the models to develop prompts for itself during inferencing in multiple iterations, allowing the model to refine the prompts. The effectiveness of this approach stands out with the team achieving the highest score in strict-matching precision.

Abdelmalak Abdelmalak (2025) primarily focused on Task A. They used SpaCy to tokenize the answers into sentences and then matched the labels based on proportional alignment with the reference data for training and development. Following this, they fine-tuned COVID-Twitter-BERT on two tasks: one to identify relevant sentences and the other to label each relevant sentence based on its perspective.

LMU Agustoslu (2025) participated only in Task B and evaluated a set of different prompting techniques for the summarization task. They achieved high performance in relevancy metrics through the use of fine-tuning and few-shot learning based approaches. Competitive performance was achieved in the factuality metrics by deploying a variant of Chain-of-thought reasoning known as SumCoT, which was designed for element extraction and text summarization tasks.

Human Analysis We conducted a thorough human analysis of the summaries by the top 5 teams based on five criteria defined by Fabbri et al. (2021). The human annotator annotates 15 summaries generated by the top 5 teams for this evaluation on a Likert scale from 1-5. These criteria are as follows:

1. **Coherence**- Is the generated summary coherently framed?
2. **Consistency**- Is the summary logically implied by the source answer?
3. **Fluency**- How well-formulated is the summary gramatically?
4. **Relevance**- Does the summary include only relevant and non-redundant information from the source answers?
5. **Coverage**- How well is the particular perspective covered in the summary?

The results of the human analysis based evaluation are given in Table 5. Based on this evaluation, we identified Team YALENLP (Jang et al., 2025) and Team KHU_LDI as consistently producing the highest quality of summaries. This observation is consistent with our evaluation using the automatic metrics where Team YALENLP (Jang et al., 2025) achieved the best scores in the relevance metrics. The high fluency and coherence scores for all teams are expected outcomes of using LLMs for generation, as these models are capable of producing high-quality, grammatically correct English text. However, relevance remains a weak point for all submissions, as the models often produce elaborate, unrelated, and irrelevant content. Consistency scores indicate how well the model follows the flow and logic of the user's answers, with Team KHU_LDI performing the best in this metric. Coverage is strong for some models, while others often miss key pieces of information, an issue that we believe can be mitigated by more effective utilization of the predicted spans as input.

7 Conclusion

This work presents an overview of the PerAnsSumm 2025 Shared Task, organized at the CL4Health Workshop 2025 which received 155 total submissions from 23 teams. The task aimed to identify and summarize perspective spans in answers in Community Question-Answering forums. Specifically, it contains two subtasks: (a) Perspective Span Identification and Classification and (b) Perspective-based Summarization. To this end, this task utilized the PUMA dataset (Naik et al., 2024) that was supplemented with a newly annotated test set for evaluation. We described relevant performance metrics for this task and provided an overview of our findings, as well as the approaches taken by the 12 teams that submitted system papers. We are optimistic that the provided resources will help foster further research toward the task of perspective-based answer summarization. To enable future work, we continue maintaining the CodaBench webpage for the Shared Task as a benchmark.

Limitations

The PerAnsSumm shared task involves generation of summaries which are evaluated automatically while presenting the leaderboard. This involves the selection of automatic metrics, which, while a strong indicator, may not be completely representative of actual summary quality. For this reason, we include a range of diverse evaluation metrics. Due to the number of participants, we conduct our human evaluation study only on the summaries generated by the top five participants which may be expanded to include all participants to determine the correlation between the human evaluation and automatic metrics in future work. Further, the wide use of LLMs in the shared task encourages us to define metrics more suited for evaluating LLM generated content in future runs of this shared task. These evaluations which were not included in the current shared task may include evaluating specifically for LLM hallucinations along with the current evaluation of factuality.

References

Abanoub Abdelmalak. 2025. Abdelmalak at peranssumm 2025: Leveraging a domain-specific bert and llama for perspective-aware healthcare answer summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Tanalp Agustoslu. 2025. Lmu at peranssumm 2025: Llama-in-the-loop at perspective-aware healthcare answer summarization task 2.2 factuality. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Anson Antony, Peter Vickers, and Suzanne Wendelken. 2025. Roux-lette @ peranssumm shared task. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Abari Bhattacharya, Rochana Chaturvedi, and Shweta Yadav. 2022. LCHQA-summ: Multi-perspective summarization of publicly sourced consumer health answers. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 23–26, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Rochana Chaturvedi, Abari Bhattacharya, and Shweta Yadav. 2024. Aspect-oriented consumer health answer summarization. *Preprint*, arXiv:2405.06295.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dongsuk Jang, Alan Li, and Arman Cohan. 2025. Yalenlp @ peranssumm 2025: Multi-perspective integration via mixture-of-agents for enhanced healthcare qa summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Jooyeon Lee, Luan Huy Pham, and Özlem Uzuner. 2025. Mnlp at peranssumm: A classifier-refiner architecture for improving the classification of consumer health user responses. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Sushvin Marimuthu and Parameswari Krishnamurthy. 2025. Ltrc-iiith at peranssumm 2025: Spansense - perspective-specific span identification and summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.

Vansh Nawander and Nerella Chaithra Reddy. 2025. Datahacks at peranssumm 2025: Lora-driven prompt engineering for perspective aware span identification and summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024a. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Tabea M. G. Pakull, Hendrik Damm, Henning Schäfer, Peter A. Horn, and Christoph M. Friedrich. 2025. Wispermed @ peranssumm 2025: Strong reasoning through structured prompting and careful answer selection enhances perspective extraction and summarization of healthcare forum threads. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kristin Qi, Youxiang Zhu, and Xiaohui Liang. 2025. Team_umb at peranssumm 2025: Enhancing perspective-aware summarization with prompt optimization and supervised fine-tuning. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rakshith R, Mohammed Sameer Khan, and Ankush Chopra. 2025. Aicoe at peranssumm 2025: An ensemble of large language models for perspective-aware healthcare answer summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Pablo Romero, Libo Ren, Lifeng Han, and Goran Nenadic. 2025. The manchester bees at peranssumm 2025: Iterative self-prompting with claude and o1 for perspective-aware healthcare answer summa. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Nadia Saeed. 2025. Medifact at peranssumm 2025: Leveraging lightweight models for perspective-specific summarization of clinical qa forums. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *Preprint*, arXiv:2406.04692.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.